Open Cloze 化した語彙穴埋め問題の運用可能性評価
BERT 予測に基づく正解語保持判定法

北川みやび[1] 竹中要一[2]
[1]関西大学大学院 総合情報学研究科 [2]関西大学 総合情報学部
{k189216, takenaka}@kansai-u.ac.jp

概要

本研究は，多肢選択式語彙穴埋め問題を Open Cloze 問題へ機械変換した際に，元問題の正解語を単一正答として維持できているとみなせる問題かどうかを，モデル予測に基づき自動判定する枠組みを提案する．BERT の masked language model（MLM）で空所の予測分布を得て，正解語と top-1 予測語の一致／不一致により妥当性を操作的に定義した．英検 2 級 30 問から 1 トークン条件で 28 問を評価し，一致は 5 問（17.9%）であった．さらに，不一致事例の一部では正解語が上位候補に含まれることが確認され，Open Cloze 化に伴う別解の扱いと採点基準設計が課題となることを指摘する．

1 はじめに

第二言語学習において語彙は基盤的要素であり，学習者は語彙量だけでなく，語の形式・意味・用法（共起，頻度，レジスター等）を含む多面的な知識を獲得する必要がある[1]．したがって，語彙知識の到達度を多面的に把握する評価手段の設計は教育実践上重要である．

語彙知識の評価方法の一つとして，多肢選択式穴埋め問題（以下，多肢選択式語彙問題）は広く用いられている[2]．多肢選択式は採点が迅速かつ安定しやすく，大規模実施に適する点が利点である[2]．

一方，選択肢を与えない自由記述式の穴埋め問題（以下，Open Cloze 問題）は，解答を選択肢から選ぶのではなく，学習者が文脈に合う語を自力で生成（想起）して空所を埋めることを要求する形式である[3]．Pino らは，Open Cloze 問題では文脈に基づく語の産出が求められるため，綴りや文体選択などを含む産出的知識も評価し得ると述べている[5]．しかし，選択肢制約を外すことで許容解が増え得るため[3]，採点時に想定外の解答を正答として扱うかの判断が必要になり，特に大規模実施では運用上の負担が大きい[2]．このような課題がある一方で，多肢選択式語彙問題の中には，選択肢を外して Open Cloze 化しても，元問題の正解語が依然として有力な正答として成立する問題が含まれる可能性がある．しかし，既存の多肢選択式語彙問題を Open Cloze として転用する場合，どの問題が「選択肢なしでも元の正解語を単一の正答として維持しやすいか」を人手で全件精査することは現実的でない．したがって，自動的にスクリーニングする枠組みが必要になる．

以上を踏まえ，本研究は，多肢選択式語彙問題を Open Cloze 問題へ変換した際に，元問題の正解語が選択肢制約なしでも正答として成立するとみなせるかを，自動的に判定する枠組みを提案する．具体的には，BERT に代表される Masked Language Model（MLM）[4] の予測を用い，正解語とモデルの top-1 予測語が一致するか(表層一致)を判定基準とする．本稿ではこの一致を，選択肢を外しても元の正解語が最有力候補として残るかを見積もるための簡便な基準として扱う．なお，不一致はこの一致基準を満たさないことを示すにとどまり，文として不自然であることを直接意味しない

2 関連研究

語彙知識は単なる形式―意味対応にとどまらず，用法・共起・レジスター等を含む多面的な知識から成る[1]．したがって，語彙評価は，どの側面を測ろうとするかに応じてタスク形式を選択する必要があり，形式の違いは測定対象の違いとして現れ得る．
本研究では 多肢選択式語彙問題(正解と誤選択肢から選択して解答する形式）と Open Cloze 問題（選択肢を与えず空所を埋める形式）を区別する[3][5]．Open Cloze 問題は，解答候補が提示されないため，学習者が文脈に合う語を自力で産出することが要求され，より挑戦的な課題として位置づけられる[3]．一方で，評価の運用面では採点の容易さが形式選択を左右し得る．多肢選択式語彙問題は，解答の判定を迅速かつ安定して行いやすく，大規模実施に適するという利点を持つ[2]．これに対し，Open Cloze 問

題では，想定外の解答を正答として扱うかどうかの判断が必要になり得るため，大規模実施では採点負担が増えやすい[2]．また，既存の公開資源は多肢選択形式を中心に整備されていることが指摘されており[3]，多肢選択式語彙問題を Open Cloze 問題へ転用できれば，既存資源を活用できる可能性がある．本研究が Open Cloze 問題を扱う理由は，選択肢による手がかりを排し，文脈に基づく語の産出を要求することで，より産出的側面を含む語彙運用を測り得る点にある[3][5]．ただし，既存の多肢選択式語彙問題を機械的に Open Cloze 問題へ変換しただけでは，元の正解語を正答として維持できるかを一貫した基準で判断しにくい．

この判断を支える実装上の要素として，本研究は Masked Language Model（MLM）を用いる．MLM は入力中の一部トークンをマスクし，その語を周辺文脈から予測する事前学習タスクとして定義される[4]．したがって，MLM が与える空所位置の予測分布は，文脈に対する補充候補の確からしさを定量化する情報として利用可能である[4]．以上を踏まえ，本研究は，多肢選択式語彙問題を Open Cloze 問題へ変換した際に，元問題の正解語が選択肢制約なしでも妥当な解として成立するとみなせるかを，モデル予測に基づき自動判定する枠組みを提案する．

3 研究手法

本章では，本研究で用いる言語モデルと，Open Cloze 化後の問題を自動判定するための提案手法を述べる．具体的には，まず 3.1 節で本研究におけるマスク言語モデル推定の基盤として BERT を概説し，`[MASK]` 位置に対する語の確率分布を得る仕組みを整理する．続いて 3.2 節で，本研究が提案する「元問題の正解語とモデルの top-1 予測語の一致／不一致に基づく妥当性判定」を定義し，その判定仕様を示す．

3.1 BERT を用いた MLM 推定

本研究では，Open Cloze 形式へ変換した空所補充文に対し，masked language model（MLM）を用いて空所に入る語の確率を推定する．MLM とは，文中の特定位置をマスク記号（例：[MASK]）で隠し，その位置に入る語を周辺文脈から予測する枠組みである[4]．本研究では MLM を実現するモデルとして BERT（Bidirectional Encoder Representations from Transformers）を用いる．BERT は Transformer に基づくエンコーダ型の言語モデルであり，左右の文脈を同時に参照して表現を構築できる点に特徴がある[4]．また，BERT は事前学習に MLM を含むため，[MASK] 位置に入る語の確率分布を推定する用途に適している[4]．

本研究で用いる BERT は，英語事前学習済みモデル bert-base-uncased である．本モデルは入力を小文字化して処理するため，推定結果として得られる [MASK] 位置の予測語も基本的に小文字で出力される．したがって，本研究では 大小文字の差は一致判定の対象外とし，正解語と予測語を小文字化したうえで比較する（例：I と i は一致として扱う）．

3.2 提案手法

具体的には，元問題の空欄に対応する位置を特定し，その位置を [MASK] に置換した英文を生成してモデルに入力する．モデル出力として得られる [MASK] 位置の語彙（トークン）ごとの確率分布から最大確率の語を top-1 予測語として採用し正解語および top-1 予測語を小文字化した表層形の完全一致により，一致／不一致を判定する（大小文字の違いは不一致としない）．判定は [MASK] 位置に入る 1 トークンを対象とするため，正解語がトークナイザにより複数トークンへ分割される問題は判定仕様と整合しない．したがって本研究では，正解語が 1 トークンとして扱える問題のみを判定対象とする．また，単複・時制などの派生形は一致扱いとしない．

本研究の判定出力は一致／不一致（Yes/No）であり，判定そのものは top-1 予測語のみに基づく．ただし，不一致となった事例の性質を分析する目的で，確率上位候補（例：top-10）も取得し，正解語が上位候補に含まれるか等を参照できるようにする．なお，この top-10 は判定規則を変更するものではなく，不一致事例分析のための補助情報として扱う．

4 実験条件

データには，英検 2 級の 2024 年度第 1 回，第 2 回，第 3 回の問題を用いた．各回から，多肢選択式の語彙穴埋め問題のうち，選択肢が単語のみで構成される問題を抽出し，各回 10 問ずつ計 30 問を対象候補とした．なお，正解語がトークナイザにより複数トークンに分割される問題は，正解語を 1 トークンとして扱う判定仕様と整合しないため除外し，最終的に 28 問を分析対象とした．

5 実験結果

5.1 一致した例

Open Cloze 形式では，正解語を 1 トークンとして扱える問題のみを対象に，モデルの top-1 予測語と正解語の一致率を算出した．その結果，28 問中 5 問で一致し（5/28，17.9%），23 問で不一致であった．一致した 5 問について，top-1 予測語の確率の分布を図 1 に示す．

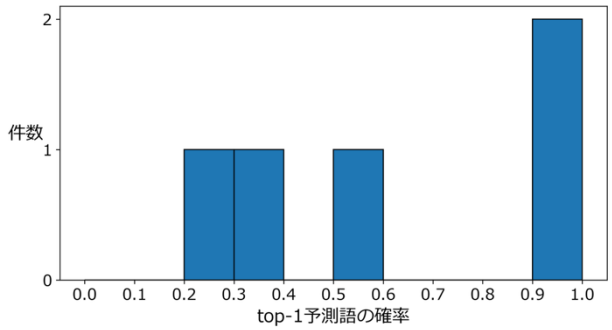

図 1 top-1 予測語の確率分布(n = 5)

5.2 不一致だった例

top-1 予測語が正解語と一致しなかった事例を示す．以下では，正解語と top-1 予測語が異なる一方で，正解語が top-10 予測内に含まれていた例を提示する．なお，本節では事例の提示に留め，原因や傾向の解釈は考察で扱う．

事例 1（minister / ministry）

本事例の正解語は minister であるが，モデルの top-1 予測語は ministry であり不一致となった（ministry: p=0.659, minister: p=0.298）．ただし minister は top-10 予測内に含まれていた．表 1 に top-10 予測語と確率を示す．

The [MASK] of education is responsible for ensuring that the country’s schools, universities and other places of learning are being managed well.

表 1 事例 1 の top-10 予測語

順位	予測語	確率
1	ministry	0.659
2	minister	0.298
3	department	0.030
4	secretary	0.004
5	directorate	0.002
6	director	0.001
7	commissioner	0.001
8	ministers	0.001
9	board	0.001
10	ministries	0.001

事例 2（cure / treat）

本事例の正解語は cure であるが，モデルの top-1 予測語は treat であり不一致となった（treat: p=0.541, cure: p=0.268）．ただし cure は top-10 予測内に含まれていた．表 2 に top-10 予測語と確率を示す．

Although medicine has made a lot of progress in the past 50 years, researchers still have not found a way to [MASK] a cold.

表 2 事例 2 の top-10 予測語

順位	予測語	確率
1	treat	0.541
2	cure	0.268
3	prevent	0.034
4	combat	0.014
5	reverse	0.012
6	heal	0.011
7	control	0.009
8	relieve	0.001
9	overcome	0.001
10	stop	0.001

6. 考察

本研究では，多肢選択式語彙問題を Open Cloze 問題へ変換した際に，元問題の正解語が正答として成立するとみなせるかを，MLM の予測に基づいて自動判定する枠組みを提案した．具体的には，元問題の正解語と MLM の top-1 予測語の一致／不一致を判定基準として用いた．正解語を 1 トークンとして扱える 28 問を対象に評価した結果，top-1 予測語と正解語が一致した問題は 5 問（5/28，17.9%）であり，不一致は 23 問（23/28，82.1%）であった．

まず，一致が得られた 5 問は，選択肢集合という制約を外した Open Cloze 問題においても，正解語がモデルの最有力候補（top-1）として選ばれる例である．したがって，本研究の一致基準（top-1 一致）に従う限り，多肢選択問題を機械的に Open Cloze へ変換しても，そのまま成立する問題が一定数存在

することが確認された．一方で一致率は 17.9%に留まっており，無条件に Open Cloze 問題に変換すれば同等の語彙問題として成立する，とは言い難い．よって，Open Cloze 化後に採用可能な問題を選別する手続きが必要になる．

次に，一致事例（n=5）における top-1 予測語確率の分布を図 1 に示す．一致した 5 問では top-1 確率が 0.25〜0.96 の範囲に分布し，0.9 以上の高確信度を示す例も含まれていた．少数例ではあるが，この結果は，モデルが高い確信度で top-1 を出す問題ほど Open Cloze 問題として安定して成立する可能性を示唆する．ただし，採用基準を確率により厳しくすると採用できる問題数は減少するため，精度と収量のトレードオフが生じる．また本研究のデータ数は限られるため，しきい値の妥当性は追加データによる検証が必要である．

一方，ただし，不一致は「本研究の一致基準（top-1 一致）を満たさない」ことを示すにとどまり，Open Cloze として成立しないことを直接意味しない．不一致事例の中には，正解語が top-10 内に含まれ，しかも上位に位置する例が存在した（例：minister/ministry，cure/treat；表 2，表 3）．これらの例は，文脈上成立し得る語が複数存在する状況では，モデルが正解語以外の候補を top-1 として選ぶことがあり得ることを示している．実際，Open Cloze 問題を対象としたデータセット研究では，各空所に対して許容解を想定し，空所が一意解に限られないことが示唆されている[3]．この非一意性は評価・採点の観点では扱いにくい一方で，学習目的であれば，複数の妥当解を比較させて「なぜその語が文脈に合うのか／どこが不自然か」を説明させる課題として活用できる可能性もある．

さらに，Open Cloze 問題は選択肢が提示されないため，学習者が文脈に基づいて語を自力で産出することを要求する点で，処理負荷の高い課題として位置づけられる[3][5]．語彙学習方略に関する研究では，意味的な関連づけや既有知識との結び付けなど，より深い処理を伴う方略の使用が語彙知識（語彙サイズ）と関連することが示されており，一方で単純反復のみでは学習成果との一貫した関連が示されにくいことが報告されている[6]．したがって，Open Cloze 問題を学習タスクとして用いる場合，単に正解語を当てさせるだけでなく，候補語の比較や理由の説明を促す設計を組み合わせることで，関連づけを伴う処理を引き出し，語彙学習に寄与する可能性がある．ただし，この点は学習方略研究の知見に基づく推論であり，本研究は Open Cloze 問題の学習効果そのものを検証していない．学習目的での有効性は今後，課題形式（多肢選択／Open Cloze）やフィードバック方法を統制した実験により検討する必要がある．

したがって，本研究のように正解語を 1 語に固定したうえで top-1 の表層一致のみを妥当性基準とする判定は，Open Cloze に内在する「解の非一意性」を意図的に切り捨てる厳格な基準になっている可能性がある．

ただし，上位候補には語彙問題として不適切な語が含まれ得るため，上位候補をそのまま許容解として扱うことは妥当でない．さらに，許容解を増やす運用を取る場合でも，想定外応答を正答として採点するかの判断が必要になり，採点ルール設計や運用コストが増大し得る．以上より，不一致事例は，単一解の厳格判定だけでは捉え切れない側面があることを示しつつも，許容解候補を設計するには候補集合の制御と，運用可能な採点方針の設計が不可欠であることを示唆する．

7. 今後の展望

今後は，まず採用基準（確率しきい値）の検証を進める必要がある．top-1 確率を用いた採用フィルタは有望である可能性がある一方，精度と収量のトレードオフを伴う．したがって，データ数を増やした上で，しきい値ごとに採用件数と一致率を整理し，運用上妥当なしきい値設計を行う．

次に，不一致事例で正解語が上位候補に残る例が確認されたことを踏まえ，許容解候補集合の構成規則を検討する必要がある．上位候補をそのまま許容解とするのは緩すぎるため，候補語の絞り込み（例：確率差，名詞・動詞などの品詞制約，語彙レベル，意味的近さ）を導入し，候補集合を制御する手続きを設計する．

さらに，候補集合が構成できたとしても，それが解として違和感がないか，あるいは学習に有益かは別問題である．今後は少数でもよいので人手評価により候補語の受容性を点検し，その上で単一解形式と複数解形式を比較して学習効果や解答行動への影響を実証的に検討する．

参考文献

[1] Norbert Schmitt. 2008. Instructed second language vocabulary learning. Language Teaching Research, 12(3):329–363.

[2] Ralph Rose. 2020. Improving te Production Efficiency and Well-formedness of Automatically-Generated Multiple-Choice Cloze Vocabulary Questions. In Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), pages 7094–7101, Marseille, France. European Language Resources Association (ELRA).

[3] Mariano Felice, Shiva Taslimipoor, Øistein E. Andersen, and Paula Buttery. 2022. CEPOC: A Cambridge Exams Publishing Open Cloze Dataset. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022), pages 4285–4290, Marseille, France. European Language Resources Association (ELRA).

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[5] Juan Pino, Michael Heilman, and Maxine Eskenazi. 2008. A Selection Strategy to Improve Cloze Question Quality. In Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains, pages 22–34, Montreal, Canada (ITS 2008 workshop).

Philippe Fournier-Viger

[6] 内田奈緒. 2021. 中高の英語学習における語彙学習方略の使用と英語学力の関連. 教育心理学研究, 69(4), 366‐381.